빅콘테스트
- 딥러닝과 시뮬레이션을 통한 야구 경기 예측

성균관대 통계학과 이경택
서울대학교 보건대학원 김강진

데이터수집

NAVER SPORTS 이용 - 크롤링

- 팀별 일별 경기기록 수집(2013~2015)
- 경기별 상세 기록 수집(선발 투수 방어율, 불펜 투수 방어율, 타자기록 등)

➡ 팀별 데이터 관리

[KIA 데이터]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	날짜	요일	득점1	득점2	팀1	팀2	지역	요일코	연속일	주말.주	원정.홈	결과	득점	실점	상대팀	home_	home_	away_	away_	home_	home_	away_	away_	defen	bul_de	attack	attack	최근te	yester
2	9|14	(일)	6	10	KIA	한화	대전	7	1	주말	원정	패	6	10	한화	5.65	NA	8.08	6.788	12	0.288	13	0.288	8.08	6.788	13	0.288	0.3	승
3	9|13	(토)	3	1	KIA	한화	대전	6	1	주말	원정	승	3	1	한화	4.7	4.14	4.8	4.228	4	0.288	9	0.288	4.8	4.228	9	0.288	0.4	패
4	9|12	(금)	4	14	KIA	삼성	대구	5	1	주말	원정	패	4	14	삼성	4.28	6.35	4.27	5.914	19	0.304	11	0.288	4.27	5.914	11	0.288	0.3	패
5	9|11	(목)	4	5	KIA	삼성	대구	4	1	평일	원정	패	4	5	삼성	5.33	4.36	6.09	4.606	13	0.303	7	0.288	6.09	4.606	7	0.288	0.3	패
6	9|10	(수)	12	6	LG	KIA	광주	3	1	평일	홈	패	6	12	LG	6.06	6.4	5.17	5.117	10	0.289	10	0.277	6.06	6.4	10	0.289	0.3	승
7	9|9	(화)	7	10	LG	KIA	광주	2	2	평일	홈	승	10	7	LG	7.76	5.09	4.16	4.629	10	0.289	13	0.277	7.76	5.09	10	0.289	0.4	패
8	9|7	(일)	3	6	KIA	NC	마산	7	1	주말	원정	패	3	6	NC	3.75	4.685	5.7	5.168	11	0.287	6	0.289	5.7	5.168	6	0.289	0.3	승
9	9|6	(토)	9	0	KIA	NC	마산	6	6	주말	원정	승	9	0	NC	5.29	4.575	3.82	4.391	6	0.287	15	0.29	3.82	4.391	15	0.29	0.3	패
10	8|31	(일)	4	3	SK	KIA	광주	7	1	주말	홈	패	3	4	SK	7.5	4.457	9	6.024	10	0.289	11	0.29	7.5	4.457	10	0.289	0.2	패
11	8|30	(토)	7	2	SK	KIA	광주	6	1	주말	홈	패	2	7	SK	5.79	NA	4.18	5.713	7	0.289	12	0.289	5.79	NA	7	0.289	0.3	승
12	8|29	(금)	9	5	KIA	롯데	사직	5	1	주말	원정	승	9	5	롯데	4.26	10.95	3.98	6.573	8	0.289	12	0.289	3.98	6.573	12	0.289	0.4	패
13	8|28	(목)	5	6	KIA	롯데	사직	4	1	평일	원정	패	5	6	롯데	5.57	4.978	6.75	5.597	11	0.289	11	0.289	6.75	5.597	11	0.289	0.4	패
14	8|27	(수)	4	6	KIA	넥센	목동	3	2	평일	원정	패	4	6	넥센	3.57	2.565	4.03	4.436	8	0.298	8	0.289	4.03	4.436	8	0.289	0.5	패
15	8|25	(월)	9	0	한화	KIA	광주	1	1	평일	홈	패	0	9	한화	5.6	5.47	5.53	NA	3	0.289	13	0.29	5.6	5.47	3	0.289	0.5	승
16	8|23	(토)	5	6	한화	KIA	광주	6	1	주말	홈	승	6	5	한화	4.15	6.387	3.12	5.436	9	0.29	12	0.289	4.15	6.387	9	0.29	0.5	패
17	8|22	(금)	2	3	KIA	LG	잠실	5	5	주말	원정	패	2	3	LG	3.75	3.19	3.57	3.337	9	0.279	8	0.291	3.57	3.337	8	0.291	0.4	패

분석방법론

야구경기 승/패 요인

야구팬들의 대화

1) 오늘 어디랑 해?

- 약팀이냐 vs 강팀이냐
- 먹고먹히는 앙숙 관계

➡ 상대팀이 중요

2) 오늘 선발 누구야?

- 야구는 투수 놀음

➡ 선발・불펜 방어율 중요

3) 최근 분위기 좋더라?

- 야구는 분위기 싸움

➡ 최근 경기 결과 중요

4) 홈경기야?

- 유난히 홈경기에 강한 팀

➡ 홈/원정 경기 중요

분석방법론

야구경기 승/패 요인

예측구간 선수 능력치/ 그때 당시에 팀 최근 승률?

분석방법론

모형설정 – 딥러닝 사용

딥러닝 : 인공신경망의 진화된 형태

팀 별 모든 경기 예측

적용변수

지역, 요일코드, 연속 경기 일수, 주말.주중 여부, 원정.홈 여부, 상대팀,

선발방어율, 불펜 방어율, 타율, 어제 경기결과, 연속 승수, 최근10경기승률

※ 선발투수방어율 : 대부분 5선발 체제로 어느 정도 추정 가능 ± 오차

※ 타자타율 : 시즌 막바지이기 때문에 평균 타율은 큰 변화 없음 ± 오차

1경기 예측 시, 2종류의 모델사용

- ① 해당 팀의 올해 전적 데이터 승리 예측 – 승리 확률
- ② 상대 전적 데이터(올해 + 작년) – 승리 확률

① 의 확률 + 가중치 ② ➡ 경기 승/패 예측

Ex) ① KIA 승리확률 0.82 (승리) / ② KIA승리 확률 0.31(패배)
=> 0.82 – 0.2 = 0.62 (승리 예측)

2014~2013년 시뮬레이션을 통해 선정

분석방법론

모형설정 – 딥러닝 사용

Bagging 기법적용

- 여러가지 모형을 결합해 안정적인 성능을 낼수 있도록 하는 기법

- 모형설명력 향상
- 독립변수에 난수추출 등 Random한 요소존재
- 시뮬레이션이 목적이기 때문에, 다수의 모형이 결합된 안정된 모형 사용
- Bootstrap시 최근의 데이터에 가중치 (야구는 분위기 싸움)

분석방법론

시뮬레이션

9월 6일 KIA vs 삼성

KIA 전체 데이터프레임

상대팀	어제결과	최근승률	..
LG	승	0.7	..
두산	패	0.8	..

KIA vs 삼성 데이터프레임

상대팀	어제결과	최근승률	..
삼성	승	0.5	..
삼성	패	0.6	..

딥러닝 모형 10회반복(Bagging)

① KIA 승리 확률 : 0.782

② KIA 승리 확률 : 0.412(가중치로 사용)

결과 예측 : KIA 승리 확률 : 0.782 – 0.2 : 0.582 (KIA 승리)

9월 8일 KIA vs NC

KIA 전체 데이터프레임

상대팀	어제결과	최근승률	..
삼성	승	0.8	..
LG	승	0.7	..

KIA vs NC데이터프레임

상대팀	어제결과	최근승률	..
NC	승	0.7	..
NC	패	0.6	..

분석방법론

알고리즘

1) 팀데이터 추출(ex : KIA)

2) 상대팀 전적 데이터 추출(ex : 삼성)

3) 딥러닝 모델 생성 – 1)의 데이터를 통한 모델링
2)의 데이터를 통한 모델링

(10회반복) Bagging 기법 적용

4) KIA가 이길 확률(1) + 가중치(2)

5) 4)의 결과가 맞다고 가정, 다음 변수(승률, 최근 경기 결과 등) 추정

5) 1) ~ 5) 반복

시뮬레이션 10회씩 반복

결과

9월30일 기준

예측

순위	팀	승	무	패	승률
1	삼성	85	0	56	0.602
2	NC	81	2	58	0.581
3	두산	78	0	63	0.553
4	넥센	77	1	63	0.549
5	한화	68	0	73	0.482
5	KIA	68	0	73	0.482
7	롯데	67	0	74	0.475
7	SK	66	2	73	0.475
9	LG	60	1	80	0.429
10	kt	52	0	89	0.3687

실제

순위	팀	승	무	패	승률
1	삼성	85	0	56	0.603
2	NC	82	2	56	0.594
3	넥센	76	1	64	0.543
4	두산	76	0	64	0.543
5	SK	68	2	71	0.489
5	한화	67	0	74	0.475
7	KIA	66	0	73	0.475
7	롯데	65	1	75	0.464
9	LG	62	2	76	0.449
10	KT	51	0	89	0.364

Summary

시뮬레이션

Deep Learning + Ensemble Method(Bagging)

최근 분위기 데로 갈 확률 + 징크스에 발목 잡힐 확률

경기 결과 예측

장점	단점
-앙상블 기법과 시뮬레이션을 통한 안정성 증대 - 두 모형을 사용함으로써 예측력 강화	- 추가적인 변수 고려 필요성 - 긴 시간의 학습과정

감사합니다